1. The research group wanted to study the diversity of plants in the Galapagos Islands. The group measured the number of different plants in each island (total of 30 islands) $Y = \text{Species}$ and measured the values of the different geographic variables for each island:

   $X_1 =$ Area - Surface area of island, hectares,
   $X_2 =$ Elevation - Elevation in m,
   $X_3 =$ Nearest- Distance to closest island, km,
   $X_4 =$ Scruz- Distance from Santa Cruz Island, km,
   $X_5 =$ Adjacent- Area of closest island, hectares.

	name	Species	Area	Elevation	Nearest	Scruz	Adjacent
1	Baltra	58	25.09	346	0.6	0.6	1.84
2	Bartolome	31	1.24	109	0.6	26.3	572.33
3	Caldwell	3	0.21	114	2.8	58.7	0.78
4	Champion	25	0.10	46	1.9	47.4	0.18
5	Coamano	2	0.05	77	1.9	1.9	903.82
6	Daphne.Major	18	0.34	119	8.0	8.0	1.84
.							
.							
29	Tortuga	16	1.24	186	6.8	50.9	17.95
30	Wolf	21	2.85	253	34.1	254.7	2.33

   The data set can be found at the file galapagos.txt.

(a) Let us assume that the random variable Y_i follows the Poisson distribution $Y_i \sim Poi(\mu_i)$. Consider modeling the expected value μ_i by the following model

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}.$$

Calculate the maximum likelihood estimate for the expected value μ_{i_*} when the explanatory variables X_1, X_2, X_3, X_4, X_5 have the values

Area	Elevation	Nearest	Scruz	Adjacent
58.27	198	1.1	88.3	0.57

Construct also 95% confidence interval for the expected value μ_{i_*}.

(2 points)

(b) Assume $Y_i \sim Poi(\mu_i)$. Let us assume that the appropriate link function is square root link $g(\mu_i) = \sqrt{\mu_i}$. Consider the following hypotheses

$$H_0 : \sqrt{\mu_i} = \beta_0 + \beta_1 x_{i1},$$
$$H_1 : \sqrt{\mu_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}.$$

Select the appropriate test statistic to test the above hypotheses. Calculate the value of the test statistic, and return it as your answer to the question.

(1 point)

(c) Assume $Y_i \sim Poi(\mu_i)$. Consider modeling the expected value μ_i by the following exponential Poisson model

$$\mu_i = e^{\beta_0} x_{i1}^{\beta_1} x_{i2}^{\beta_2} x_{i3}^{\beta_3} x_{i4}^{\beta_4} x_{i5}^{\beta_5}.$$

Create 80 % prediction interval for new observation Y_f, when the explanatory variables X_1, X_2, X_3, X_4, X_5 have the values

Area	Elevation	Nearest	Scruz	Adjacent
58.27	198	1.1	88.3	0.57

(2 points)

(d) Let us assume that the random variable Y_i follows the negative binomial distribution $Y_i \sim NegBin(\mu_i, \theta)$. Consider the model

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_5 x_{i5}.$$

Calculate the fitted value $\hat{\mu}_1$ for the island $i = 1 =$ Baltra.

(1 point)

2. Consider the dataset in the file chromoabnormal.txt:

```
  cells  ca doseamt doserate
1   47800  25       1     0.10
2  190700 102       1     0.25
3  225800 149       1     0.50
4  232900 160       1     1.00
5  123800  75       1     1.50
6  149100 100       1     2.00
.
27   14400 206     5.0     4.00

An experiment was conducted to determine the effect of gamma radiation
on the numbers of chromosomal abnormalities observed

A data frame with 27 observations on the following 4 variables.
cells - Number of cells
ca - Number of chromosomal abnormalities
doseamt - amount of dose in Grays
doserate - rate of dose in Grays/hour

Purott R. and Reeder E. (1976)
The effect of changes in dose rate on the yield of chromosome
aberrations in human lymphocytes exposed to gamma radiation.
Mutation Research. 35, 437-444.
```

Focus in the study is to model how the ratio between variables Y=ca and t=cells

$$Z = \frac{Y}{t} = \frac{\text{ca}}{\text{cells}}$$

depends on the explanatory variables X_1=doseamt and X_2=doserate. Let us also first assume that $Y_i \sim Poi(\mu_i)$.

(a) Consider the log link model with interaction term

$$\mathcal{M}_{12}: \quad \log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}.$$

Calculate the maximum likelihood estimate for the expected value μ_{i_*} when $x_{i_*1} = 4$, $x_{i_*2} = 0.75$, and $t_{i_*} = 64070$.

(2 points)

(b) Calculate the maximum likelihood prediction for the ratio

$$\frac{Y_f}{t_f}$$

when $x_{f1} = 4$, $x_{f2} = 0.75$. Also, create suitable prediction intervals for the ratio $\frac{Y_f}{t_f}$.

(2 points)

(c) Assume that $\mathrm{Var}(Y_i) = \phi\mu_i$. Test at 5% significance level, is the explanatory variable X_2=doserate statistically significant variable in the model

$$\mathcal{M}_{12}: \quad \log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2},$$

Calculate the value of the test statistic.

(1 point)

(d) Consider the model

$$\mathcal{M}_{12}: \quad \log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}.$$

Under which distribution, the model $\mathcal{M}_{12}$ fits best on data in your opinion?

i. Y_i follows Poisson distribution with the variance $\mathrm{Var}(Y_i) = \mu_i$,
ii. Y_i follows quasi-Poisson distribution with the variance $\mathrm{Var}(Y_i) = \phi\mu_i$,
iii. Y_i follows negative binomial distribution $Y_i \sim NegBin(\mu_i, \theta)$.

Report to which findings you have based your decision. (1 point)

3. (a) In case of generalized linear model $g(\mu_i) = \beta_0 + \beta_1 x_i$, the maximum likelihood estimates for the parameters β_0 and β_1 are $\hat{\beta}_0 = 1$ and $\hat{\beta}_1 = 0.5$. At the value $x_i = 5$, calculate the maximum likelihood estimate of μ_i, when the model is

   i. $Y_i \sim Poi(\mu_i)$ and $\log(\mu_i) = \beta_0 + \beta_1 x_i$,

   ii. $Y_i \sim Poi(\mu_i)$ and $\sqrt{\mu_i} = \beta_0 + \beta_1 x_i$,

   iii. $Y_i \sim Poi(\mu_i)$ and $\log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_i$, where $t_i = 10$.

   iv.

$$P(Y_i = 0) = \theta_i + (1 - \theta_i)e^{-\mu_i},$$
$$P(Y_i = y_i) = (1 - \theta_i)\frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, \quad y_i = 1, 2, 3, \ldots$$
$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip},$$

   where the maximum likelihood estimate for the parameter θ_i is $\hat{\theta}_i = 0.25$.

(2 points)

(b) Let $Y_i \sim Poi(\mu_i)$. Then the probability density function of the random variable Y_i is

$$f(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}.$$

Show that the distribution of the random variable Y_i belongs to the Exponential Family of Distributions.

(2 points)

(c) In generalized linear models, the likelihood equations can written in form

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0, \qquad j = 0, 1, 2 \ldots p.$$

Consider now the most simplest Poisson model with the identity link function

$$Y_i \sim Poi(\mu_i),$$
$$\mu_i = \eta_i = \beta_0.$$

What kind of more simplified form the likelihood equations have in this case? That is, what form $\frac{\partial l(\beta_0)}{\partial \beta_0}$ has in the simplest Poisson model? Also, show by the likelihood equations that the maximum likelihood estimate of β_0 is the sample mean $\hat{\beta}_0 = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

(2 points)